Acoustic Noise Reduction by Two Dimensional Spectral Smoothing and Enhancement.

AUTHOR(S):

Ariki, Yasuo; Kajimoto, Kazuo; Sakai, Toshiyuki

STUDIA PHONOLOGICA XXI (1987)

Acoustic Noise Reduction by Two Dimensional Spectral Smoothing and Enhancement

Yasuo ARIKI, Kazuo KAJIMOTO and
Toshiyuki SAKAI

ABSTRACT

This paper proposes an acoustic noise reduction method which can recover and enhance speech formant structure as well as diffuse and suppress the noise component. A Gaussian filter is applied over a time sequence of the spectral envelope to diffuse the noise component and to recover the formant structure. The filter is called TDSS according to its function of Two Dimensional Spectral (a time sequence of the spectral envelope) Smoothing. After the application of the TDSS operator, Non-linear Spectral Amplitude Transformation (NSAT) is carried out to further suppress the noise component and enhance the formant structure. The effectiveness of this noise reduction method is shown by auditory testing and word recognition experiment.

I. INTRODUCTION

Acoustic noises such as car klaxons, jet stream of airplanes and so on sometimes disturb smooth communication through a telephone line. The noises also decrease the speech quality recorded on a tape, or recognition accuracy of a speech recognizer. To facilitate smooth communication and high recognition accuracy under a noisy environment, a noise reduction filter is required which can pass and enhance the speech as well as attenuate the noise. The purpose of this study is to develop such a filter by incorporating perceptual properties into conventional noise reduction approaches.

The conventional approaches for noise reduction may be classified into two groups. [1] One approach tries to reduce noise component by using noise properties. The typical example is a frequency subtraction method which reduces the noise by subtracting its spectrum from the noisy speech. [2] It makes good use of the noise spectrum and the non-correlation property between the speech and the noise. This approach can indeed reduce the noise, but, it is difficult for it to increase the speech intelligibility because it does not utilize speech properties. The other approach tries to extract a speech component based on speech properties and

Kazuo KAJIMOTO (梶木一夫): Yasuo ARIKI (有本康雄): Assistant Professor, Department of Information Science, Kyoto University.
Toshiyuki SAKAI (坂井利之): Professor, Department of Information Science, Kyoto University.

structure. The typical example is a comb filtering method which can extract the speech spectral amplitude only at the harmonics of the pitch frequency. [3] This approach can indeed increase the speech intelligibility, but causes so called "musical noise" which sounds like background music. A common lack in both approaches is an active estimation of formant structure which is key information for our perception of speech.

We propose, in this paper, a new approach to noise reduction, based on perceptual properties such as formant structure. It recovers and enhances the formant structure to increase the intelligibility. It also diffuses and suppresses the noise component to reduce the noisiness and to prevent the musical noise. These two operations necessary for noise reduction are integrated in our approach.

We have to solve two problems in developing the operations required for our noise reduction. The first is how to estimate the formant structure. When noise is superimposed on speech, the formant structure becomes unclear and is sometimes destroyed, then the speech intelligibility decreases. To estimate formants from noisy speech, a simple method like formant extraction based on frame by frame spectral analysis may be applied. This method, however, causes "musical noise" so that the speech quality decreases rapidly as the SN ratio decreases. [4] This musical noise is caused by discontinuity of formants between frames because of their inaccurate estimation under the noise. To estimate continuous formants accurately, inter-frame information should be used as well as intra-frame information. In this paper, we introduce a perceptual model to estimate continuous formants by using both kinds of information. We call the formant estimation based on the perceptual model a time-frequency Two Dimensional Spectral Smoothing (TDSS) operation. [5] The TDSS is a smoothing operation which applies a two-dimensional normal distribution function to a time sequence of the spectral envelope. It can diffuse the noise component by smoothing the spectral envelope and also recover the continuous formant structure by referring to the time sequence of formants. In this sense, the TDSS operation is effective for vowel recovery in noisy speech.

The last problem is how to enhance the formant structure. As the noise is superimposed on speech, the noise component is added on to the speech component. The formant structure, peak and valley of spectrum, becomes unclear so that the intelligibility decreases. To increase the intelligibility, noise suppression and formant structure enhancement should be simultaneously carried out. [6] The Non-linear Spectral Amplitude Transformation (NSAT), proposed in this paper, is an operation which can suppress the noise component and enhance the formant peaks simultaneously on the spectral domain by non-linear function.

These operations, TDSS and NSAT, have theoretically and experimentally proven to be superior to the conventional spectral subtraction method by the auditory test and the recognition test.

II. Noise Reduction by A Cepstrum Analysis and Synthesis

A. *Cepstrum Analysis and Synthesis*

Speech wave forms are usually analyzed by parametric or non-parametric methods. The parametric analysis extracts the parameters on the basis of speech generation models like an all pole model. The typical example is an LPC analysis or an LPC-cepstrum analysis. [7] On the other hand, the non-parametric analysis extracts the parameters without the speech generation model. The typical case is an FFT-cepstrum analysis. [8] Our primary goal is to investigate what kinds of speech information are most influenced by the noise, rather than to clarify the contribution of the speech generation model to the noise reduction. We, therefore, employ non-parametric analysis, in particular, the FFT-cepstrum analysis to establish the noise reduction system.

Fig. 1 is the block diagram of the FFT-cepstrum analysis and synthesis. Input speech is at first analyzed by a short-time FFT, and the phase and amplitude spectrum are obtained. The amplitude spectrum is transformed into the cepstrum by an application of a logarithm function and an inverse FFT (IFFT). The pitch information is included in the cepstrum at the higher order (higher cepstral component). The spectral envelope is contained in the cepstrum at the lower order (lower cepstral component). Through the application of an FFT and an exponential function to the lower cepstral component, the spectral envelope is obtained. At this stage, three kinds of information are produced by the FFT-cepstrum; phase, higher cepstral component and spectral envelope. The left part in Fig. 1 corresponds to the analysis system and the right part to the synthesis system of the FFT-cepstrum analysis and synthesis.

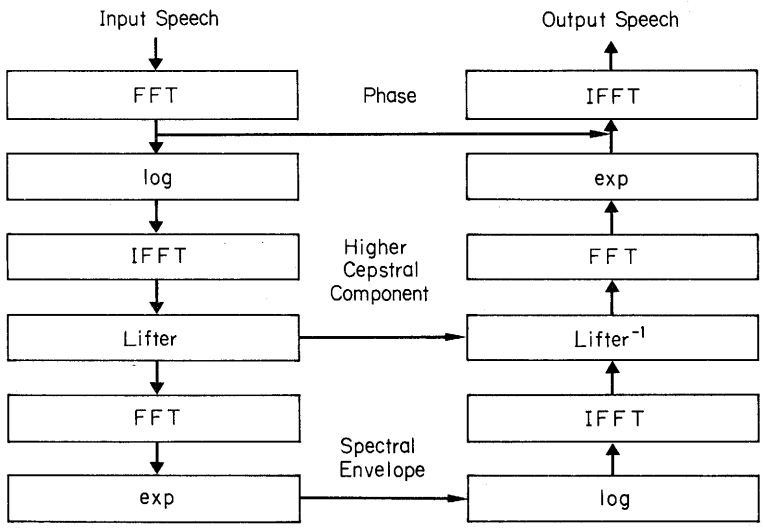


Fig. 1. Block diagram of an FFT-cepstrum analysis and synthesis.

B. Listening Tests for Noise Reduction Effects

To establish the noise reduction system, we have to clarify which information is most influenced by the noise, or on which information the noise reduction is most effective among the three. It is on this information that we can concentrate our effort to reduce the noise. Fig. 2 shows the system to produce synthesized speech for listening tests to investigate the effect of the noise reduction on the three kinds of information. The original speech is analyzed by the FFT-cepstrum and the three kinds of information are obtained. On the other hand, white noise is generated and superimposed on the original speech. The noisy speech is also analyzed by FFT-cepstrum in the same way as the original speech. As a result, three pairs are obtained: phase, higher cepstral component and spectral envelope. By selecting one from each pair, eight combinations are obtained. The FFT-cepstrum synthesizer produces these eight kinds of speeches from the combination. The condition of the FFT-cepstrum analysis and the speech materials are mentioned in Table 1.

Paired comparison listening tests were carried out. From the eight synthesized speeches, three groups of four speeches are formed by stressing on one of the three kinds of information as shown in Table 2: (a) for phase, (b) for higher cepstral component and (c) for spectral envelope. Each group includes the original speech, the noisy speech and the other two speeches which keep only one of the three kinds of information noisy or original. The four synthesized speeches in each group were presented for the paired comparison experiment in which the preferences were expressed on a scale of 7 points as below.

(3) : I prefer i to j strongly.
(2) : I prefer i to j moderately.
(1) : I prefer i to j slightly.
(0) : No preference.

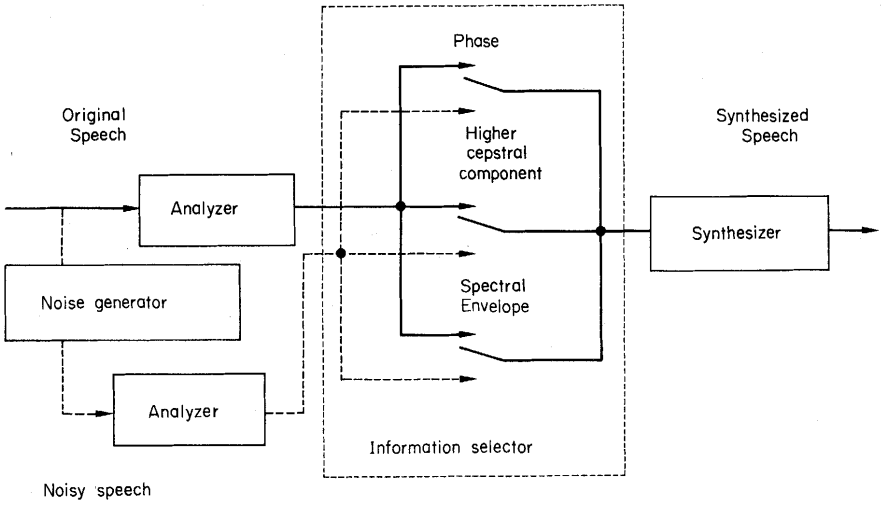


Fig. 2. System to produce the synthesized speech for listening tests.

Table 1. Condition of the FFT-cepstrum analysis for listening tests

condition	sampling frequency	10 KHz
	sampling accuracy	16 bit/point
	frame length	25.6 ms
	frame period	25.6 ms
	window	hamming window
	lifter	$1 \quad (0 \leqq x \leqq 30)$ $\cos((x-30)\pi/20) \quad (30 < x < 40)$ $0 \quad (40 \leqq x \leqq 255)$
material	speech	five Japanese vowels /a//i//u//e//o/ continuously spoken by one adult male within two seconds.
	noise	gaussian white noise, SN ratio is −10 dB: averaged power of speech to noise

Table 2. Combination of three kinds of information for listening tests
(O: noise free vowel, N: noisy vowel)

(a) Phase

	phase	higher cepstral component	spectral envelope
speech A	O	O	O
speech B	N	N	N
speech C	N	O	O
speech D	O	N	N

(b) Higher cepstral component

	phase	higher cepstral component	spectral envelope
speech A	O	O	O
speech B	N	N	N
speech E	O	N	O
speech F	N	O	N

(c) Spectral envelope

	phase	higher cepstral component	spectral envelope
speech A	O	O	O
speech B	N	N	N
speech G	O	O	N
speech H	N	N	O

(−1) : I prefer j to i slightly.
(−2) : I prefer j to i moderately.
(−3) : I prefer j to i strongly.

where i and j are the names of the stimuli and the stimulus i is presented before j. The number of subjects for the listening tests is nine for experiment (a) and (b), and seven for experiment (c). The results of the listening test were evaluated by a kind of variance analysis, the Scheffe method. [9]

C. Results of Listening Tests

The Scheffe method analyzes the preference score of the 7-point scale by five factors: main effects of stimuli, individual preference of main effects, combination

Table 3. Variance analysis of listening tests
main* indiv: individual preference of main effects
order* indiv: individual preference of order effects
df: degrees of freedom
significance: at the 1% level

(a) Phase

	sum of squares	df	variance	F ratio	significance
main effect	364.361	3	121.454	223.404	4.079*
main* indiv	24.389	24	1.016	1.896	2.068
combination effect	3.861	3	1.287	2.367	4.079
order effect	8.333	1	8.333	15.328	7.008*
order* indiv	13.000	8	1.625	2.989	2.779*
error	38.056	70	0.544		
total	452.000	108			

(b) Higher cepstral component

	sum of squares	df	variance	F ratio	significance
main effect	392.306	3	130.769	166.853	4.079*
main* indiv	27.694	24	1.154	1.472	2.068
combination effect	1.139	3	0.380	0.484	4.079
order effect	0.593	1	0.593	0.756	7.008
order* indiv	5.407	8	0.676	0.862	2.779
error	54.861	70	0.784		
total	482.000	108			

(c) Spectral envelope

	sum of squares	df	variance	F ratio	significance
main effect	220.250	3	74.417	114.834	4.171*
main* indiv	58.250	18	3.236	5.062	2.228*
combination effect	0.393	3	0.131	0.205	4.171
order effect	0.298	1	0.298	0.466	7.125
order* indiv	3.286	6	0.584	0.857	3.159
error	34.524	54	0.639		
total	317.000	84			

effects of stimuli, order effects of presentation, and individual preference of order effects. Table 3 shows the results of the variance analysis for respective groups in Table 2. The symbol "*" indicates an F ratio greater than significance at the 1% level, referring to the F-table. From these tables, it is concluded that the main effects are significant; therefore, the four speeches in each group have a significant difference. To clarify the difference among the four speeches in each group, the values of the main effect of each speech were estimated as shown in Table 4. If the difference between the values of the main effect of two speeches is greater than the "yardstick" at the 1% level, then a significant difference is found between the two speeches. From Table 4, the following are summarized.

(1) The noise reduction processing on the phase information is not prospective, from Table 4(a), for the following two reasons.
   (a) There is no significant difference at the 1% level between the noisy speech B (−1.181) and the speech D (−0.986) which is noise free only on the phase information.
   (b) The difference between the speech C (0.667) and the speech A (1.500) is less than that between the speech D (−0.986) and the speech A. This means that the noise reduction on other than the phase information is more effective.

(2) The noise reduction processing on the higher cepstral component is not prospective, from Table 4(b), for two reasons.
   (a) There is no significant difference at the 1% level between the noisy speech

Table 4. Values of main effects

(a) Phase

speech A	1.500
speech B	−1.181
speech C	0.667
speech D	−0.986

Significance at 1% level is 0.295

(b) Higher cepstral component

speech A	1.639
speech B	−1.111
speech E	0.569
speech F	−1.097

Significance at 1% level is 0.425

(c) Spectral envelope

speech A	1.643
speech B	−0.679
speech G	−0.875
speech H	−0.089

Significance at 1% level is 0.393

B (−1.111) and the speech F (−1.097) which is noise free only on the higher cepstral component.

(b) The difference between the speech E (0.569) and the speech A (1.639) is less than that between the speech F (−1.097) and the speech A. This means that the noise reduction on other than the higher cepstral component is more effective.

(3) The noise reduction processing on the spectral envelope is fairly prospective, form Table 4(c), for three reasons.

(a) A significant difference is found at the 1% level between the noisy speech B (−0.679) and the speech H (−0.089) which is noise free only on the spectral envelope.

(b) The difference between the speech G (−0.875) and the speech A (1.643) is greater than that between the speech H (−0.089) and the speech A. This means that the noise reduction on the spectral envelope is more effective than the noise reduction on the other information.

(c) There is no significant difference at the 1% level between the noisy speech B (−0.679) and the speech G (−0.875) which is noisy only on the spectral envelope. This means that the noise reduction on other than the spectral envelope is not effective.

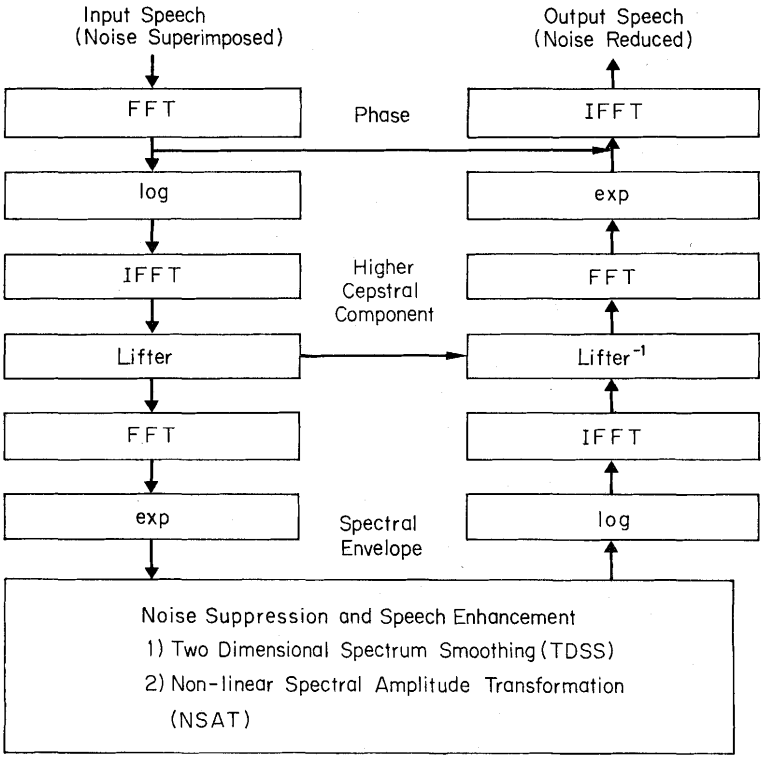


Fig. 3. Block diagram of the noise reduction system.

D. Structure of a Noise Reduction System

From the discussion in II.C, it is clear that the noise reduction on the spectral envelope is most effective, even if the other information is not processed. According to this conclusion, we constructed a noise reduction system on the basis of the FFT-cepstrum analysis and synthesis system shown in Fig. 1. Fig. 3 is the block diagram of a system in which the noise is reduced only on the spectral envelope by two methods. These two methods for noise reduction on the spectral envelope are described in sections IV and V. Before describing them, we introduce a perceptual model which is the basic concept for these noise reduction methods.

III. Perceptual Model for the Noise Reduction

A. Hypothesis of a Perceptual Model

The human auditory system can perceive speech even in an acoustically noisy environment; this is commonly known as the cocktail party effect. For noisy speech perception, we assume that formant frequencies might be perceived on the basis of a perceptual model which can diffuse the environmental noise and recover the speech formant structure. If such a kind of perceptual model can be constructed mathematically and applied to noisy speech, an accurate extraction of the formant frequencies may be expected. Here, we assume a perceptual model as follows:

"*Given an acoustic stimulus at a certain point on a time-frequency surface, the stimulus is perceived even at the neighboring points around the stimulus by a magnitude weighted by the normal distribution function.*"

According to this perceptual model, the perceived value at a certain point on the time-frequency surface is computed by summing the magnitudes of the stimuli around the point, weighted by the normal distribution function. This indicates that the perceived value is technically computed by a convolution of the normal distribution function and the input stimuli on the time-frequency surface. We can show this perceptual model's validity for both a critical band and temporal masking effectiveness as follows.

B. Validity of the Model for a Critical Band

A critical band is defined as follows: [10]

(1) If the band width of the input stimulus is less than that of the critical band, the perceived loudness is constant. Otherwise, if it exceeds the critical band width, the perceived loudness increases. This is shown in Fig. 4(d) where the perceived loudness is indicated by P and the band width of the input stimulus by W.

(2) The critical band width increases as its central frequency increases.

Here, we explain (1) by the perceptual model we have assumed. For simplicity in the analysis, we use a uniform distribution function as the perceptual model instead of the normal distribution function.

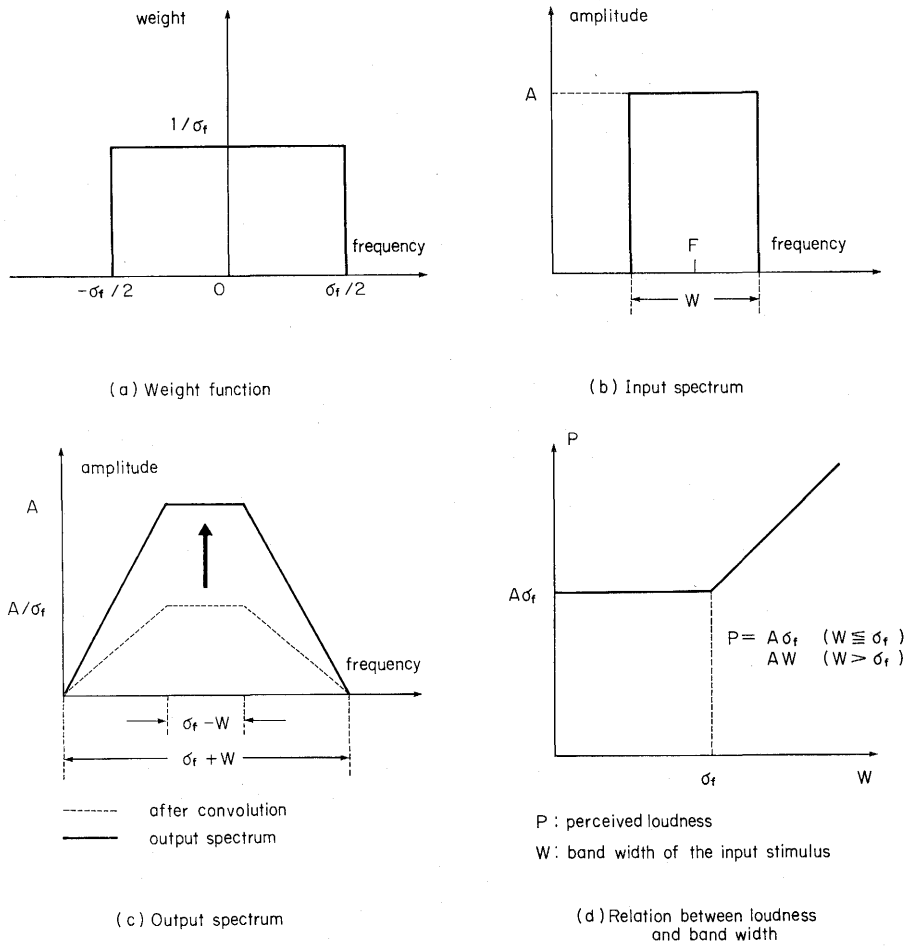


Fig. 4. Critical band.

We denote the band width of the uniform distribution function by σ_f, and its weight by $1/\sigma_f$ as shown in Fig. 4(a). If the band width of the input spectrum is W, the central frequency F, and the maximum amplitude A, as shown in Fig. 4(b), the convolution of the weight function and the input spectrum becomes a trapezoid shown by the dotted line in Fig. 4(c). Here, we assume that the maximum amplitude of the output spectrum is equal to that of the input spectrum. The solid line in Fig. 4(c) shows the result of this transformation. The perceived loudness is computed as the area of the trapezoid by integrating the output spectrum. As a result, the perceived loudness P is formulated by the following expression:

$$P=\begin{cases} A\sigma_f & (W\leq\sigma_f) \\ AW & (W>\sigma_f) \end{cases} \tag{1}$$

The relation between the perceived loudness P and the band width W of the input stimulus is shown in Fig. 4(d). The detailed derivation of expression (1) is shown in Appendix A.

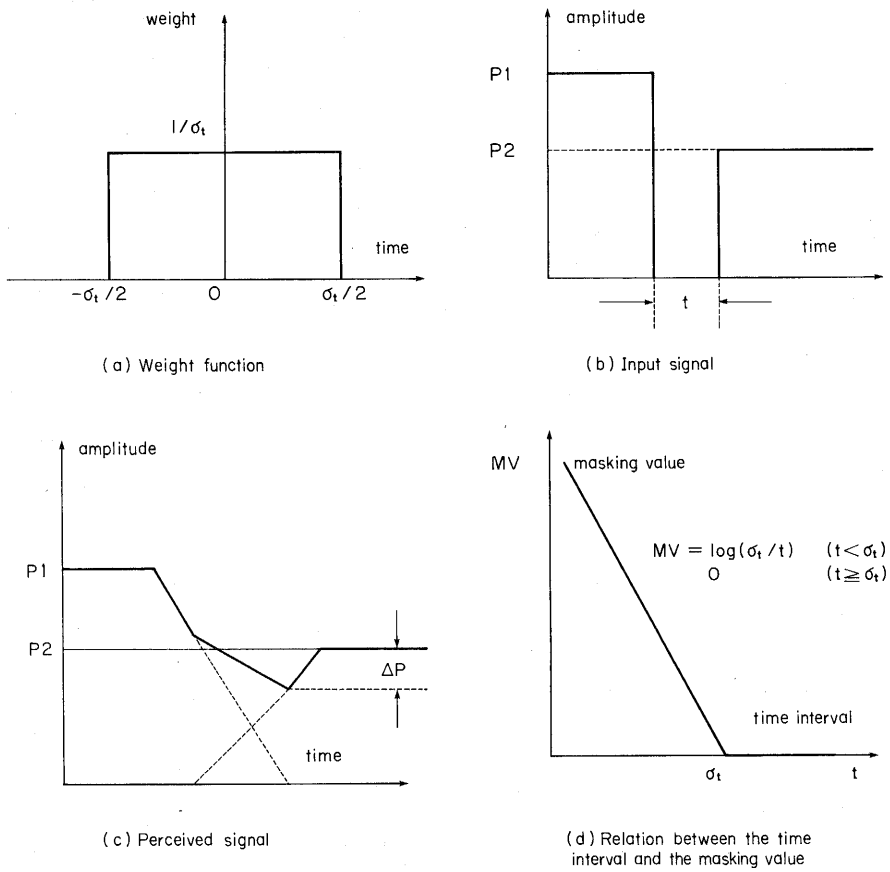
C. Validity of the Model for Temporal Masking

Temporal masking is defined as follows: [11]

(1) If two acoustic stimuli are presented in succession at a time interval t and the loudness of the first stimulus is greater than the second, the second stimulus is masked by the first stimulus. This is called a forward masking.

(2) A masking value is defined as the difference between minimum audible pressures with and without the masking stimulus. This is shown in Fig. 5(d) where the masking value MV decreases as the time interval increases, and finally becomes zero after a certain time σ_t.

Here, we explain this temporal masking by using the uniform distribution function in the same way as the critical band. For input acoustic signals with amplitudes $P1$, $P2$ and a time interval t as shown in Fig. 5(b), the weight function with time duration σ_t and weight $1/\sigma_t$ as shown in Fig. 5(a) is convoluted. Fig. 5(c) is the result of the convolution. We assume that the amplitude difference between the minimum value and the second stimulus is equal to ΔP. As a result, the masking value MV is formulated by the following expression:



(a) Weight function

(b) Input signal

(c) Perceived signal

(d) Relation between the time interval and the masking value

Fig. 5. Temporal masking effect.

$$MV=\begin{cases}\log\ (\sigma_t/t) & (t<\sigma_t)\\ 0 & (t\geq\sigma_t)\end{cases} \tag{2}$$

The relation between the time interval t and masking value MV is shown in Fig. 5(d). The detailed derivation of expression (2) is shown in Appendix B.

The normal distribution function can show the same effect for the critical band and the temporal masking effect. This has been confirmed by the computer simulation.

IV. Two Dimensional Spectral Smoothing (TDSS)

A. *Derivation of a TDSS Operation from the Perceptual Model*

According to our perceptual model, the perceived value at a certain point on the time-frequency surface is computed by convoluting the normal distribution function (NDF) to the input stimuli. The explanation of the critical band and the temporal masking described in the previous section was based on one-dimensional convolution along a frequency axis and a time axis independently. These two one-dimensional convolutions, however, can be integrated into one two-dimensional convolution of the NDF to the input stimuli on the time-frequency surface. We call this two-dimensional convolution a Two Dimensional Spectral Smoothing (TDSS) operation, and the two-dimensional NDF a TDSS operator.

The input stimuli on the time-frequency surface is a time sequence of the spectral envelopes analyzed by the FFT-cepstrum from a digitized speech wave form. We call, henceforth, these input stimuli the two dimensional spectral envelope. On this two dimensional spectral envelope, the TDSS operator is convoluted. The TDSS operator has the property of smoothing the two dimensional spectral envelope due to the low pass filter by the NDF convolution. It is for this reason that we adopt the name TDSS. Fig. 6 shows the TDSS operator $w(i\Delta f, j\Delta t)$ which can be formalized as the following expression:

$$w(i\Delta f, j\Delta t)=A\exp\left[-\frac{1}{2}\left\{\frac{(i\Delta f)^2}{\sigma_f{}^2}+\frac{(j\Delta t)^2}{\sigma_f{}^2}\right\}\right] \tag{3}$$

$$A=\left\{\sum_i\sum_j \exp\left[-\frac{1}{2}\left\{\frac{(i\Delta f)^2}{\sigma_f{}^2}+\frac{(j\Delta t)^2}{\sigma_t{}^2}\right\}\right]\right\}^{-1}$$

where i and j are the locations of the mel-frequency axis and the time axis, respectively. The σ_f and σ_t are the standard deviations of the normal distribution function and are used to control the range to be weighted. The Δf and Δt are the observation intervals on the mel-frequency and on the time, respectively. The condition for the application of the TDSS operator is the same as shown in Table 1 for litening tests, except for the frame interval from 25.6 ms to 3.2 ms. The Δt, therefore, is 3.2 ms. The mel-frequency is the log scale frequency on which the human auditory system is based. The linear frequency 1000 Hz corresponds to 1000 mel-frequency. The spectral envelope with the linear frequency f is converted to that with the mel-frequency m by the following expression:

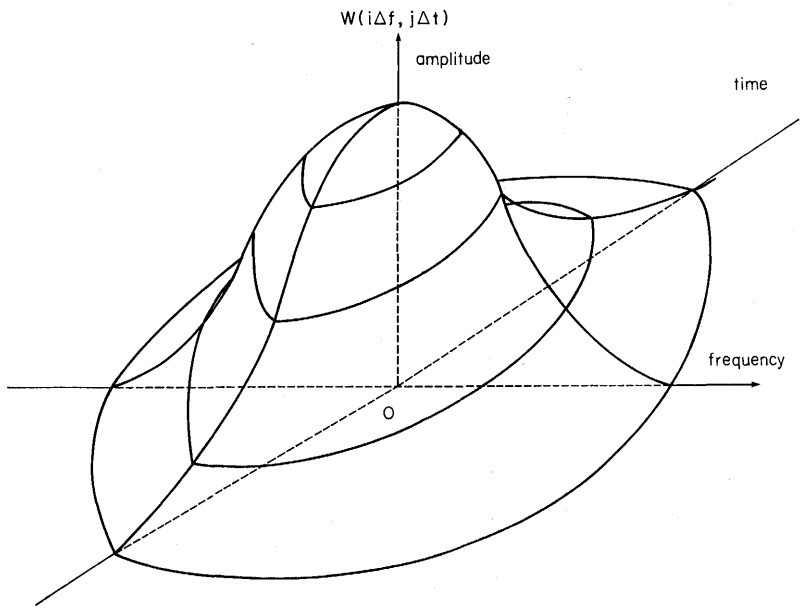


Fig. 6. TDSS operator.

$$m=1000 \log_2 (1+f/1000) \tag{4}$$

Since the sampling frequency is 10 KHz and the frame length is 25.6 ms, the spectral envelope from 0 Hz to 5,000 Hz is represented in terms of 129 frequency sampling points (f_i, $0 \leqq i \leqq 128$). The linear frequency 5,000 Hz corresponds to the mel-frequency 2585 mel according to expression (4), therefore, the mel-frequency interval Δf becomes 20.2 mel (2585 mel/128 intervals). The linear frequency f corresponding to every 20.2 mel-frequency is computed by the inverse expression of (4). The spectral envelope with mel-frequency is then obtained by interpolating the two spectral amplitudes at the linear frequencies f_i and f_i+1 ($0 \leqq i \leqq 128$) near the computed linear frequency f.

B. Effect of the TDSS Operation

Noise causes the following phenomena on the spectral envelope:

(1) Appearance of isolated peaks.

(2) Destruction of continuous peaks over a time axis.

Continuous peaks become strong cues for the speech perception because they are very likely to correspond to the formant sequence. We call, henceforth, these continuous peaks the formant ridges. Since the TDSS operation is a gaussian low-pass filter, rapid change is removed and slow change is enhanced on the two dimensional spectral envelope. Therefore, the isolated peaks which have rapid changes are removed by the TDSS operation. The formant ridges are recovered by the relative enhancement of the TDSS operator on the slow changes on the time and frequency axis. For the above reasons, the effects of the TDSS operation may be summarized as follows:

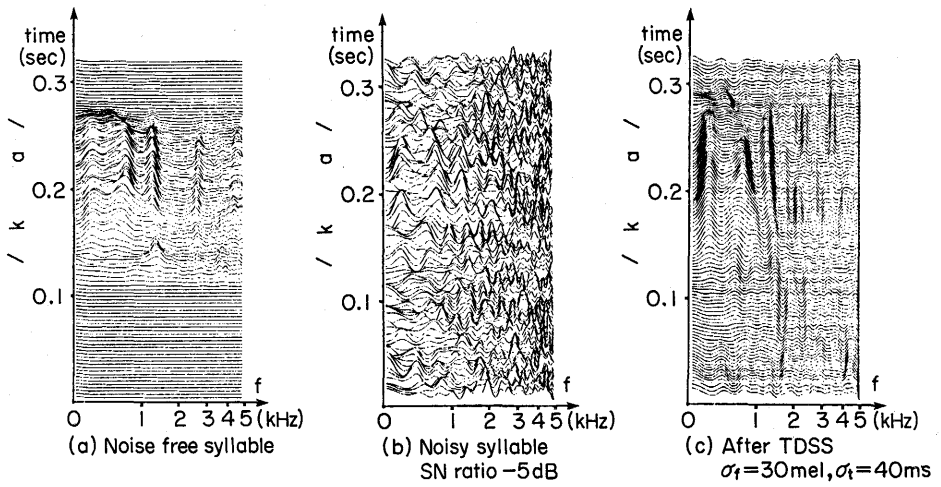


Fig. 7. Effect of the TDSS operation.

(1) It diffuses the noise component observed as isolated peaks.

(2) It recovers the speech component observed as formant ridges.

Fig. 7 shows the effects of the TDSS operation. The speech data is the Japanese syllable /ka/. Fig. 7(a) is the noise free original syllable. Fig. 7(b) is the noisy syllable with an SN ratio of −5 dB. Fig. 7(c) shows the result after the TDSS operation with $\sigma_f=30$ mel, $\sigma_t=40$ ms. It seems that the formant structure is enhanced visually after the TDSS operation on the two dimensional spectral envelope and thus it can be used to recover the vowels from the noisy speech.

C. Effect of σ_f and σ_t

To investigate the effect of the standard deviations σ_f and σ_t of the NDF in the TDSS operator, three cases are analyzed by changing their values. In each case, the range of the convolution is $-101 \text{ mel} \leqq i\Delta f \leqq 101$ mel in the mel-frequency and $-32 \text{ ms} \leqq j\Delta t \leqq 32$ ms in time. Since the mel-frequency interval Δf is 20.2 mel and the frame interval Δt is 3.2 ms, the range of i and j are $-5 \leqq i \leqq 5$ and $-10 \leqq j \leqq 10$. Fig. 8 shows the three shapes of the TDSS operator produced by changing σ_f and σ_t. In this figure, the value at $i=0$, $j=0$ is set to 100, and the values of the weight greater than 10 are shown.

(1) Case 1: $\sigma_f=30$ mel, $\sigma_t=40$ ms

Fig. 8(a) is the shape of the TDSS operator. In this case, the weight decreases gradually along the time axis and is cut down at $j=\pm 10$. The weight along the frequency axis rapidly decreases until $i=\pm 3$. This indicates that the convolution range on the time axis is 67.2 ms (3.2 ms × 21 frames) which is long enough to recover one vowel because the averaged duration of vowels is 70 ms when spoken at normal speed. The convolution range in the frequency axis is 141.4 mel (20.2 mel × 7 points) which corresponds to about 118 Hz in the band width at the central frequency 500 Hz. It is about twice as large

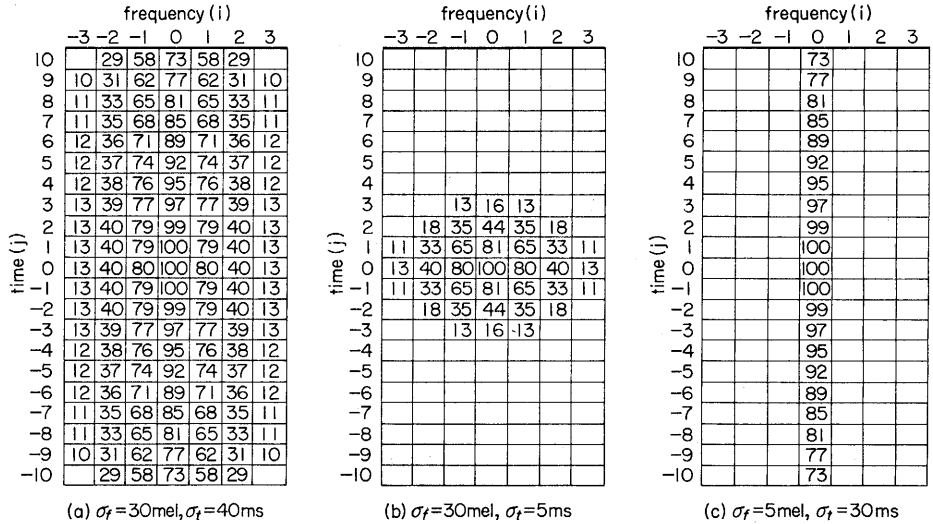


Fig. 8. Figure of the TDSS operator.

as the band width 50 Hz at the first formant (500 Hz).

(2) Case 2: $\sigma_f=30$ mel, $\sigma_t=5$ ms

Fig. 8(b) is the shape of the TDSS operator. In this case, the weight decreases rapidly along both axes and the meaningful weight range is $-3\leqq i\leqq 3$ and $-3\leqq j\leqq 3$. The convolution range on the time axis is 22.4 ms (3.2 ms$\times$7 frame). As it is too short, vowels are not recovered as shown in Fig. 9(a).

(3) Case 3: $\sigma_f=5$ mel, $\sigma_t=40$ ms

Fig. 8(c) is the shape of the TDSS operator. In this case, the weight decreases rapidly along the frequency axis till $i=\pm 1$. The convolution range on the frequency axis is 20.2 mel which corresponds to about 21 Hz in the band width at the central frequency 500 Hz. As it is shorter than the band width of the

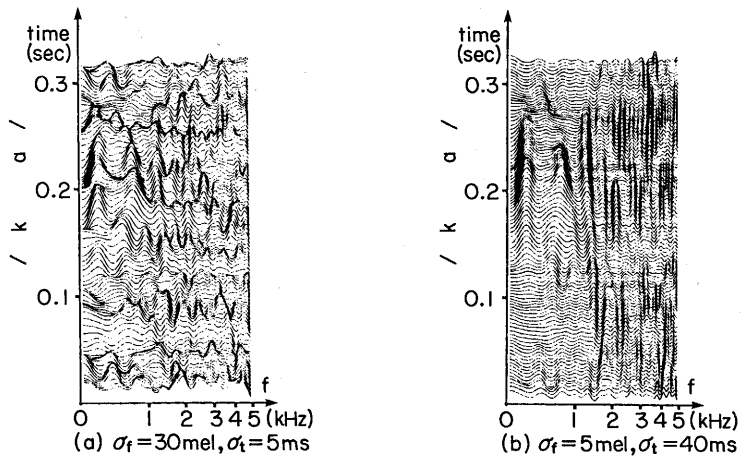


Fig. 9. Effect of the standard deviation σf, σt for the TDSS operator.

first formant, false formant ridges appear as shown in Fig. 9(b).

As a conclusion, it is reasonable that the convolution range is 67.2 ms ($-10 \leqq j \leqq 10$) on the time axis and 141.4 mel ($-3 \leqq i \leqq 3$) on the frequency axis. The standard deviations such that σ_f is 30 mel and σ_t is 40 ms realize this reasonable convolution range.

V. Speech Enhancement by a Non-Linear Spectral Amplitude Transformation (NSAT)

A. Derivation of an NSAT Operation

Noise causes the following phenomena on the spectral envelope as well as those described in IV.B.

(1) The noise component is uniformly added to the speech component.

(2) The difference between peaks and valleys is lowered.

In order to make the speech clear after the TDSS operation, the noise component must be suppressed and the peaks must be enhanced simultaneously on the spectral envelope. The NSAT operation shown in Fig. 10 is devised for these two kinds of spectral amplitude transformations. The NSAT operation consists of two linear transformations with different slopes α and β. The line with slope α, which we call an α-line, suppresses the noise component by setting α to be less than 1.

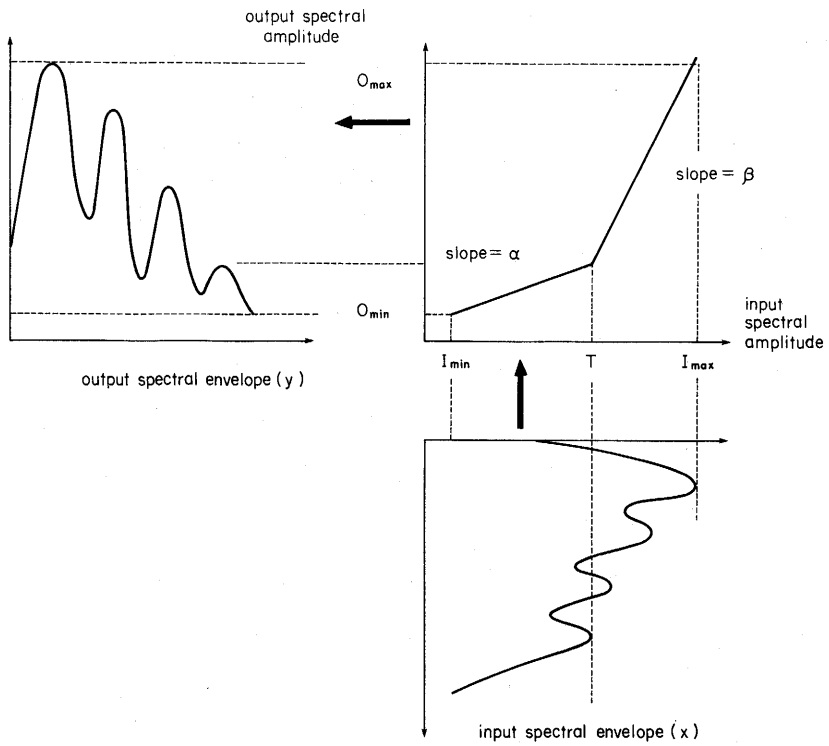


Fig. 10. NSAT operation.

The line with slope β, which we call a β-line, enhances the formant ridge by setting β to be greater than 1. The NSAT operation is formalized as follows:

$$\begin{aligned} y&=\alpha(x-I_{min})+O_{min} \qquad (x\leqq T) \\ y&=\beta(x-T)+\alpha(T-I_{min})+O_{min} \qquad (x>T) \end{aligned} \tag{5}$$

where I_{min} and O_{min} are the minimum value of the input spectral amplitude x and the output spectral amplitude y, respectively. T is the threshold for suppression and enhancement.

B. Parameters of the NSAT Operation

The NSAT operation is carried out frame by frame. The threshold T is dynamically adapted to the noise component of the input spectral amplitude at each frame. As the noise component of the input spectral amplitude, we select the amplitude of the fourth local peak on the spectral envelope after the TDSS operation in order to enhance the lower three formant peaks and suppress the higher frequency component in the vowel segment.

At present, the slope α is fixed to 0.001, and β is computed by the following expression:

$$\beta=\frac{(O_{max}-O_{min})-\alpha(T-I_{min})}{I_{max}-T} \tag{6}$$

where I_{max} and O_{max} are the maximum values of the input spectral amplitude x and the output spectral amplitude y, respectively. When T increases, the slope β increases so that the formant peaks become sharper. The slope β should also increase as the frequency increases in order to enhance the high frequency component up to the third formant for intelligibility. This is expressed as follows:

$$\beta=\left(\frac{f}{F_{max}}\right)\beta_0 \tag{7}$$

where f is the frequency, F_{max} is the maximum value of the frequency, and β_0 is a basic slope obtained from expression (6).

In silent segments, the threshold T should be adapted to the maximum value of the input spectral amplitude in order to suppress the entire spectrum of the segment. For this adaptation, the segmentation of noisy speech into vowel and silent segments is required.

C. Segmentation of Noisy Speech

The difference between the maximum amplitude and the minimum amplitude within each frame after the TDSS operation is effective as the parameter for this segmentation. If the value of the segmentation parameter is greater than a certain threshold, the frame is regarded as a vowel frame, because formant ridges are recovered on vowels by the TDSS operation. Otherwise it is a silent frame because the noise is diffused on the silence by the TDSS operation. The consonant segment can be automatically regarded as about 10 frames proceeding the vowel segment. On the consonant segment, the NSAT operation works in the same manner as on the vowel segment.

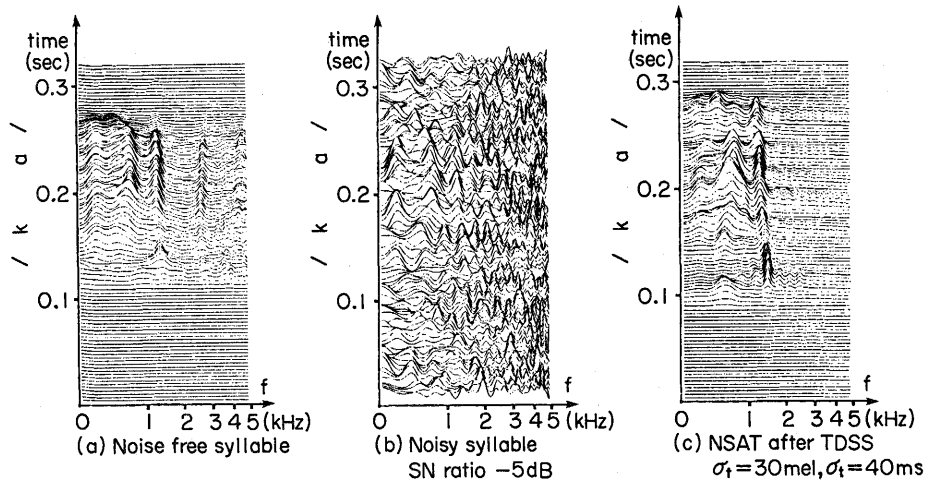


Fig. 11. Final result of the noise reduction.

D. Effect of the NSAT Operation

In the NSAT operation, if α equals 0 and β equals 1, then the noise component is simply subtracted. This transformation is well known as the spectral subtraction method. In this sense, the NSAT operation includes the spectral subtraction method and is superior to it because the NSAT operation can enhance the speech formant structure by setting the parameter β to be greater than 1. The combination of the TDSS operation and the NSAT operation produces the following effects:

(1) Isolated noises are diffused by the TDSS operation.
(2) Formant ridges are recovered by the TDSS operation.
(3) The noise component is suppressed by the α-line of the NSAT operation.
(4) Formant ridges are enhanced by the β-line of the NSAT operation.

The spectral subtraction method is equal to the special case of the NSAT operation with $\alpha=0$, $\beta=1$ in (3). In this sense, the TDSS and NSAT operations are superior to the spectral subtraction method. Fig. 11 shows the two dimensional spectral envelope of a Japanese syllable /ka/ after the TDSS and NSAT operations. In this figure, the spectral envelope at the silent segments is completely suppressed and the formant ridges are recovered and enhanced without the isolated noises.

VI. SUBJECTIVE EVALUATION EXPERIMENT

A. Preference Score by a Paired Comparison Listening Test

We evaluated our noise reduction method by the paired comparison listening test described in II.B and II.C. The compared speeches are the noisy speech, the noise reduced speech by the NSAT after TDSS, and the noise reduced speech by the simple spectral subtraction method. The SN ratio is changed in five steps: ∞, 10, 5, 0, -5 dB. The speech material is one Japanese short sentence /koNnitʃiwa/

Table 5. Value of the main effect of the preference for three speeches with five types of SN ratio

processing SN ration (dB)	∞	10	5	0	−5
unprocessed	1.629	0.352	0.007	−0.364	−1.144
spectral subtraction	0.865	−0.916	—	—	—
NSAT after TDSS	0.277	0.019	−0.091	−0.152	−0.492

(significance is 0.190 at the 1% level)

spoken by an adult male. The condition for the noise reduction is the same as that described in sections IV and V. The standard deviation of the TDSS operator is $\sigma_f=30$ mel and $\sigma_t=40$ ms. In the spectral subtraction, the noise spectrum is estimated at a certain silent segment and is fixed without adaptation. The SN ratio is changed in only two steps, ∞ and 10 dB, in the spectral subtraction. In total, 12 speeches were presented for the paired comparison tests. The number of subjects was eleven.

The result of the experiment is analyzed by the Scheffe method described in II.C. Table 5 shows the estimated value of the main effects of the preference for the three kinds of stimuli. Fig. 12 shows the graph of Table 5. From the graph, the following are summarized:

(1) SN ratio=∞ dB

The noise reduced speech by the spectral subtraction is preferred to that by the NSAT after TDSS, because the estimated noise spectrum is zero in the spectral subtraction, in principle, at ∞ dB SN ratio.

(2) SN ratio=10 dB

The noisy speech is significantly preferred to the noise reduced speech by the NSAT after TDSS, because the 10 dB noise is not so strong for the human auditory system. The spectral subtraction decreases its preference due to the

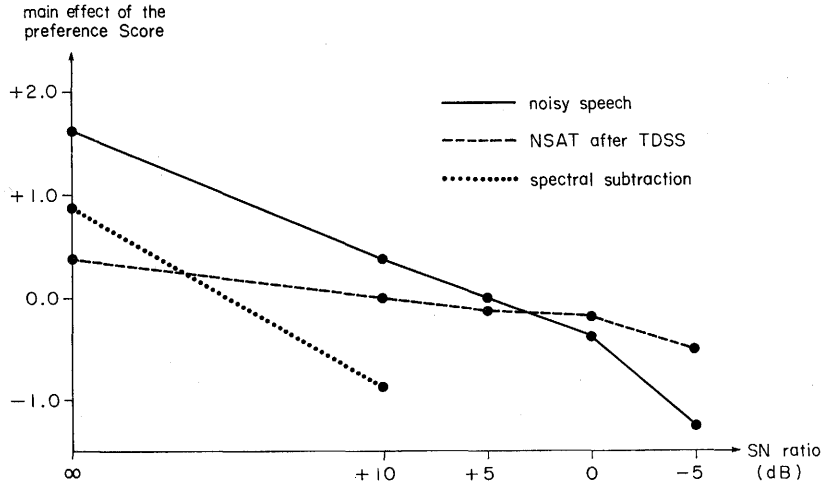


Fig. 12. Main effect of the preference.

Table 6. Value of the main effects of the preference for three processed speeches

processing	main effect
(1) NSAT only	−0.222
(2) NSAT after TDSS (σ_f=30 mel, σ_t=5 ms)	−0.093
(3) NSAT after TDSS (σ_f=30 mel, σ_t=40 ms)	0.315

(significance is 0.250 at the 1% level)

noise estimation errors.

(3) SN ratio=5 dB

As there is no significant difference between the noisy speech and the noise reduced speech, they are equal in preference.

(4) SN ratio=0 dB, −5 dB

The noise reduced speech by the NSAT after TDSS is significantly preferred to the noisy speech. This may be attributed to the disturbance of listening by the strong noise below 0 dB SN ratio. From the above investigation, it will be said that the noise reduction by the NSAT after TDSS can improve the speech quality at lower than 0 dB SN ratio in comparison with the unprocessed noisy speech.

B. Evaluation of the TDSS Operation

The TDSS operation was evaluated by a paired comparison listening test. The speech materials were five Japanese vowels /a/ /i/ /u/ /e/ /o/ spoken by an adult male with white noise of −10 dB SN ratio superimposed on them. The compared speeches for the listening test are the noise reduced speech by the three methods: (1) NSAT operation only, (2) NSAT after TDSS (σ_f=30 mel, σ_t=5 ms), and (3) NSAT after TDSS (σ_f=30 mel, σ_t=40 ms). The number of subjects was nine.

Table 6 shows the main effect of the preference estimated by the Scheffe method from the paired comparison experiment. In Table 6, there is a significant difference between speech (1) and (3), and between speech (2) and (3). This indicates that the noise reduction by the NSAT after TDSS (σ_f=30 mel, σ_t=40 ms) is effective for the human auditory system. The low preference of speech (1) and (2) may be attributed to the musical noise produced by the NSAT operation with inadequate TDSS parameters or without the TDSS operation because the diffusion of isolated noises and the recovery of formant ridges are insufficiently achieved. Inversely this means that the TDSS with adequate parameters has the effect of preventing the musical noise produced by speech enhancement such as the NSAT operation.

VII. RECOGNITION EXPERIMENT

A. Vowel Recognition Under Noise

Vowel recognition was performed to evaluate the noise reduction method by the

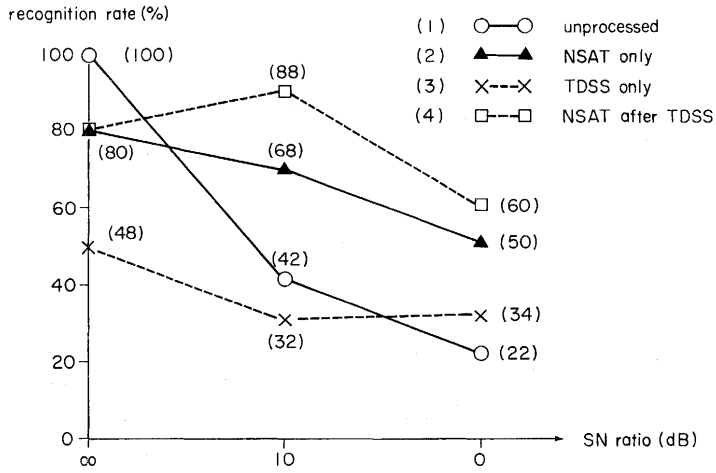


Fig. 13. Result of vowel recognition.

TDSS and NSAT. The speech materials were five Japanese vowels spoken by one adult male with noises of 10 dB and 0 dB SN ratios superimposed on them. The following four types of processing were performed on these speech materials;

(1) No noise reduction (unprocessed)
(2) Noise reduction by the NSAT only
(3) Noise reduction by the TDSS only
(4) Noise reduction by the NSAT after TDSS.

These four speech data with three SN ratio (∞, 10, 0 dB) were presented as the input patterns for recognition. Four kinds of standard patterns were produced by applying the above four processes to the noise free (∞ dB) vowels. Each of them was used to recognize the input patterns processed by the corresponding noise reduction. The input pattern was the FFT-cepstrum parameters from 0 to 13-th order, and the matching distance was the city block distance.

The result is shown in Fig. 13. The following are summarized from this figure.

(1) The NSAT operation shows a high recognition rate for noisy speech compared to the unprocessed speech, because the NSAT can enhance the formant peaks.
(2) The NSAT operation after TDSS shows the highest recognition rate for the noisy speech, because the TDSS can diffuse isolated noises and recover the formant ridges.
(3) The TDSS operation itself has the lowest, but constant recognition rate. This is because the TDSS lowers the formant ridges a little as well as diffusing the isolated noises regardless of the SN ratio.

B. Word Recognition Under Noise

Word recognition was carried out to evaluate the noise reduction method by the TDSS and NSAT. The speech materials were Japanese 30 city names spoken by one adult male with noises of 10 dB and 0 dB SN ratios superimposed on them. The following three processes were performed on these speech materials:

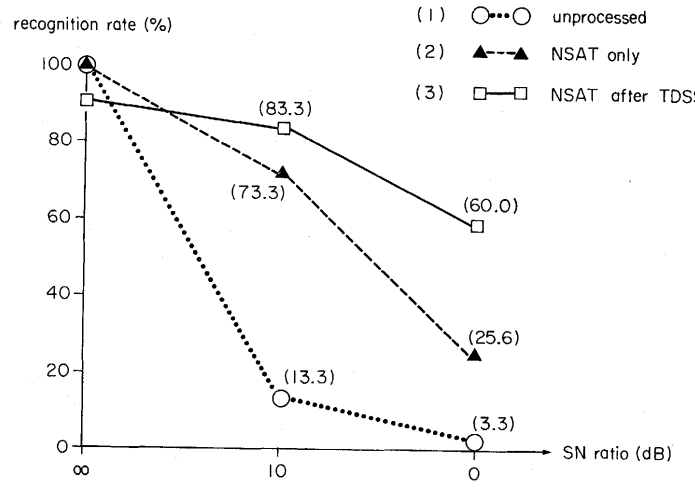


Fig. 14. Result of word recognition.

(1) No noise reduction (unprocessed)
(2) Noise reduction by the NSAT only
(3) Noise reduction by the NSAT after TDSS

These three speech data with three SN ratios (∞, 10, 0 dB) were presented as the input patterns for recognition. Three kinds of standard patterns were produced by applying the above three processes to the noise free (∞ dB) spoken words. Each of them was used to recognize the input patterns processed by the corresponding noise reduction. The input pattern was the FFT-cepstrum parameters from 0 to 13-th order, and the matching distance was the city block distance. Word segmentation is performed by visible inspection except for process (3), NSAT after TDSS. The recognition algorithm was DP-matching with end point free. The result is shown in Fig. 14. The following are summarized from the figure.

(1) At ∞ dB SN ratio, noise reduction by the NSAT after TDSS decreases its recognition rate by 10% because the TDSS lowers the formant ridges as well as diffusing isolated noises.
(2) At 10 dB SN ratio, the NSAT operation increases the recognition rate by 60% compared to the unprocessed speech. In addition, the NSAT after TDSS increases the recognition rate by 10%.
(3) At 0 dB SN ratio, the NSAT operation increases recognition rate by 22.3% compared to the unprocessed speech. In addition, the NSAT after TDSS increases recognition rate by 34.4%.

From these data, it may be said that the NSAT operation after TDSS is effective for word recognition under noise.

VIII. CONCLUSION

We have described a noise reduction method which diffuses and suppresses the noise component, and recovers and enhances the formant information.

On the FFT-cepstrum analysis and synthesis system, it has been shown by listening tests that the noise reduction on the spectral envelope is most effective for the human auditory system, in comparison with the noise reduction on the phase information or the higher cepstral component.

The perceptual model, which can explain the critical band in frequency and the temporal masking in time, was proposed on the two-dimensional spectral envelope. This model utilizes inter-frame information over time as well as intra-frame information for frequency. Based on this model, the TDSS operation was designed which can diffuse the noise component and recover the formant structure by referring to the long range in time and to the middle range of frequency.

To enhance the speech formant structure and suppress the noise component simultaneously, the NSAT operation was proposed. This is applied to the two-dimensional spectral envelope smoothed by the TDSS operation. It has been clarified that the NSAT operation includes the spectral subtraction method theoretically and experimentally.

Our new noise reduction method was evaluated by subjective preference score and computer recognition experiments. The subjective preference score revealed that the noise reduction by the NSAT after TDSS is most effective for strong noisy speech with SN ratio less than 0 dB. The computer recognition experiments also revealed that the noise reduction with the NSAT alone is effective at high SN ratios like 10 dB, but when the SN ratio decreases under 0 dB, the NSAT operation after TDSS becomes most effective. From these experiments, it may be concluded that the NSAT operation is effective at high SN ratios due to its local peak enhancement, and that the NSAT operation after TDSS is effective at the low SN ratio due to its smoothing and recovering operation.

The process of smoothing and enhancing the speech wave form reminds us of the Difference of Gaussian (DOG) filter which can extract the desired visual information by changing the standard deviation σ in the gaussian filter. [12]

Our future work will be as follows:

(1) Consonant recovery and enhancement must be achieved by extending the TDSS operator to consonants. Adaptation of the standard deviation σ to the consonant may be required in the same way as the DOG filter. According to our informal experiments, the formant transition still remains under the noise so that formant transition recovery and enhancement will be most plausible.

(2) Applicability of our noise reduction method must be clarified to several different kinds of noises such as car or helicopter noises as well as white noise.

(3) Recognition of phonemes under noisy environments must be studied for continuous noisy speech recognition.

APPENDIX A

We define the following denotations.

$f(x)$: input spectrum
$w(x)$: uniform distribution function
$g(x)$: result after convolution of $f(x)$ and $w(x)$
$max\{k(x)\}$: maximum value of $k(x)$

from the above definition output spectrum $h(x)$ is obtained as follows:

$$h(x)=a\cdot g(x) \tag{A.1}$$
$$g(x)=w(x)\otimes f(x) \tag{A.2}$$
$$a=\text{mas}\{f(x)\}/\max\{g(x)\} \tag{A.3}$$

$f(x)$ and $w(x)$ are expressed by a step function $u(x)$ as follows:

$$f(x)=A\cdot[u\{x-(F-w/2)\}-u\{x-(F+w/2)\}] \tag{A.4}$$
$$w(x)=\sigma_f{}^{-1}\cdot[u\{x-(-\sigma_f/2)\}-u\{x-(\sigma_f/2)\}] \tag{A.5}$$

The perceived loudness P is computed by integrating the output spectrum $h(x)$

$$P=\int h(x)dx \tag{A.6}$$

Here, we define $F(s)$, $W(s)$, $G(s)$, $H(s)$ as the Laplace transformation of $f(x)$, $w(x)$, $g(x)$ and $h(x)$ respectively. Using the Laplace transformation, (A.1), (A.2), (A.4) and (A.5) are expressed as follows:

$$H(x)=a\cdot G(s) \tag{A.7}$$
$$G(s)=W(s)\cdot F(s) \tag{A.8}$$
$$F(s)=A\cdot[\exp\{-(F-W/2)s\}-\exp\{-(F+W/2)s\}]/s \tag{A.9}$$
$$W(s)=\sigma_f{}^{-1}\cdot[\exp\{-(-\sigma_f/2)s\}-\exp\{-(\sigma_f/2)s\}]/s \tag{A.10}$$

Therefore, $G(s)$ is obtained by multiplying (A.9) and (A.10)

$$\begin{aligned}G(s)=A\cdot\sigma_f{}^{-1}\cdot[&\exp\{-(F-W/2-\sigma_f/2)s\}\\&-\exp\{-(F-W/2+\sigma_f/2)s\}-\exp\{-(F+W/2-\sigma_f/2)s\}\\&+\exp\{-(F+W/2+\sigma_f/2)s\}](/s^2\end{aligned}$$

There are two possibilities for the phase value. One is:

$$F-W/2-\sigma_f/2<F-W/2+\sigma_f/2<F+W/2+\sigma_f/2$$

The other is:

$$F-W/2-\sigma_f/2<F+W/2-\sigma_f/2<F+W/2+\sigma_f/2$$

Three cases are analyzed according to the phase value.

(a) $F+W/2-\sigma_f/2<F/W/2+\sigma_f/2$ $(E<\sigma_f)$

(I) $x<F-W/2-\sigma_f/2$

$$g(x)=0$$

(II) $F-W/2-\sigma_f/2\leqq x<F+W/2-\sigma_f/2$

$$g(x)=A\cdot\sigma_f{}^{-1}\cdot\{x-(F-W/2-\sigma_f/2)\}$$

(III) $F+W/2-\sigma_f/2\leqq x<F-W/2+\sigma_f/2$

$$g(x)=A\cdot\sigma_f{}^{-1}\cdot W$$

(IV) $F-W/2+\sigma_f/2\leqq x<F+W/2+\sigma_f/2$

$$g(x)=-A\cdot\sigma_f{}^{-1}\{x-(F+W/2+\sigma_f/2)\}$$

(V) $F+W/2+\sigma_f/2 \leqq x$

$$g(x)=0$$

In this case, $\max\{f(x)\}=A$, $\max\{g(x)\}=A\cdot\sigma_f{}^{-1}\cdot W$, then $a=\sigma_f/W$. Consequently,

$$P=\int a\cdot g(x)dx=A\cdot\sigma_f$$

(b) $F-W/2+\sigma_f/2=F+W/2-\sigma_f/2$ $(W=\sigma_f)$

(I) $x<F-\sigma_f$

$$g(x)=0$$

(II) $F-\sigma_f \leqq x<F$

$$g(x)=A\cdot\sigma_f{}^{-1}\cdot\{x-(F-\sigma_f)\}$$

(III) $x=F$

$$g(x)=A$$

(IV) $F<x<F+\sigma_f$

$$g(x)=-A\cdot\sigma_f{}^{-1}\cdot\{x-(F+\sigma_f)\}$$

(V) $F+\sigma_f \leqq x$

$$g(x)=0$$

In this case, $\max\{g(x)\}=A$, then $a=1$. Consequently,

$$P=A\cdot\sigma_f$$

(c) $F+W/2-\sigma_f/2>F-W/2+\sigma_f/2$ $(W>\sigma_f)$

(I) $x<F-W/2-\sigma_f/2$

$$g(x)=0$$

(II) $F-W/2-\sigma_f/2 \leqq x<F-W/2+\sigma_f/2$

$$g(x)=A\cdot\sigma_f{}^{-1}\cdot\{x-(F-W/2-\sigma_f/2)\}$$

(III) $F-W/2+\sigma_f/2 \leqq x<F+W/2-\sigma_f/2$

$$g(x)=A$$

(IV) $F+W/2-\sigma_f/2 \leqq x<F+W/2+\sigma_f/2$

$$g(x)=-A\cdot\sigma_f{}^{-1}\cdot\{x-(F+W/2+\sigma_f/2)\}$$

(V) $F+W/2+\sigma_f/2 \leqq x$

$$g(x)=0$$

In this case, $\max\{g(x)\}=A$, then $a=1$. Consequently,

$$P=A\cdot W$$

Appendix B

(a) In the case of $t \leqq \sigma_t$

Fig. B1 shows the case of $t \geqq \sigma_t$. In this case, the second stimulus value $P2$ is equal to δ which is the minimum auditory pressure without the masking stimulus. Therefore, the masking value MV is:

$$MV=\log P2/\delta=\log\delta/\delta=0$$

(b) In the case of $t<\sigma_t$

Fig. B2 shows the case lf $t<\sigma_t$. In this case, the second stimulus value $P2$ is computed as follows:

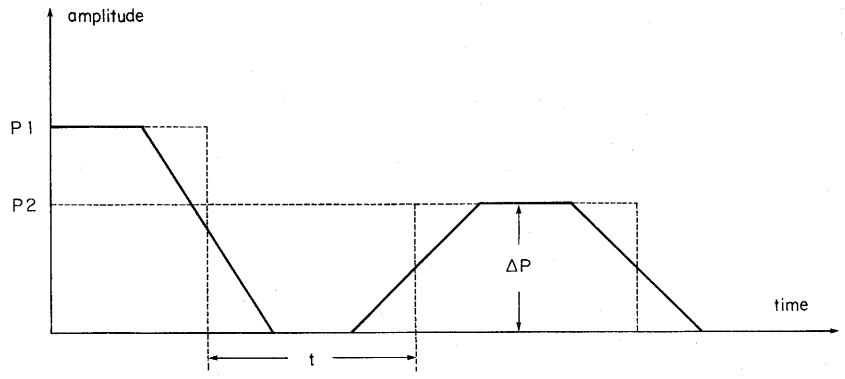


Fig. B1. In a case of $\mathbf{t} \geqq \sigma_t$.

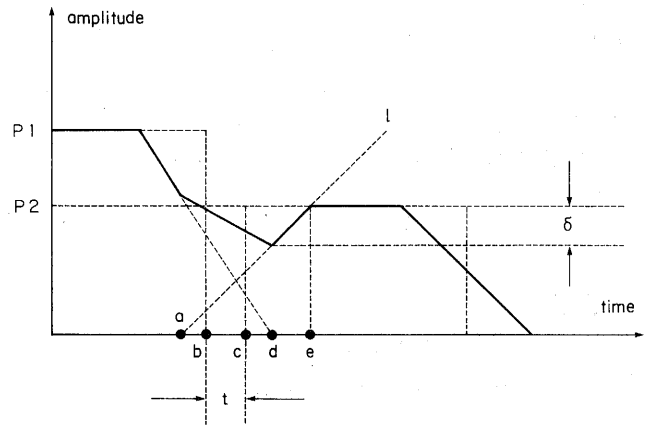


Fig. B2. In a case of $t < \sigma_t$.

$$P2 = ae \cdot slant \tag{B.1}$$

$$ae = \sigma_t \tag{B.2}$$

where *slant* is the slant of line l and is expressed as:

$$slant = \delta / de \tag{B.3}$$

$$de = ce - cd \tag{B.4}$$

where $ce = bd = \sigma_t/2$, then

$$de = ce - cd = bd - cd = bc = t \tag{B.5}$$

Consequently, according to (B.1), (B.2), (B.3) and (B.4),

$$P2 = \sigma_t \cdot \delta / t \tag{B.6}$$

Masking value MV is:

$$MV = \log P2 / = \log \sigma_t / t \tag{B.7}$$

REFERENCES

1) J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.
2) S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech & Signal Process.*, vol. ASSP-29, pp. 113–120, 1979.

3) R. H. Frazier, S. Samsam, L. D. Braida and A. V. Oppenheim, "Enhancement of speech by adaptive filtering," in *Proc. IEEE Int. Conf. Acoust., Speech & Signal Process.*, pp. 251–253, 1976.
4) M. M. Sondhi, C. E. Schmidt and L. R. Rabiner, "Improving the quality of a noisy speech signal," *Bell Syst. Tech. J.*, vol. 60, no. 8, pp. 1847–1859, 1981.
5) K. Kajimoto, Y. Ariki and T. Sakai, "Acoustic noise reduction by two dimensional spectral smoothing and spectral amplitude transformation," in *Proc. IEEE Int. Conf. Acoust. speech & Signal Process.*, pp. 97–100, 1986.
6) J. E. Porter and S. F. Boll, "Optimal estimation for spectral restoration of noisy speech," in *Proc. IEEE Int. Conf. Acoust. Speech, & Signal Process.*, 18A.2, 1984.
7) J. D. Markel and A. H. Gray, Jr, "Linear prediction of speech," *Springer-Verlag*, 1976.
8) A. V. Oppenheim, "Speech analysis-synthesis system based on homomorphic filtering," *J. Acoust. Soc. Am.*, vol. 45, pp. 458–465, 1969.
9) H. Scheffe, "An analysis of variance for paired comparisons," *Am. Statis. Assoc., J.*, vol. 47, pp. 381–400, 1952.
10) E. Zwicker, "Subdivision of the audible frequency range into critical bands," *J. Acoust. Soc. Am.*, vol. 33, no. 2, pp. 248, 1961.
11) L. L. Elliot, "Backward and forward masking of probe tones of different frequencies," *J. Acoust. Soc. Am.*, vol. 34, no. 8, pp. 1116–1117, 1962.
12) D. Marr and E. Hildreth, "Theory of edge detection," *Proc. R. Soc. Lond, B*, vol. 207, pp. 187–217, 1980.

(Aug. 31, 1987, received)